The Interplay between Entropy and Variational Distance Part I: Basic Concepts and Bounds

Siu-Wai Ho and Raymond W. Yeung

Abstract

For two probability distributions with finite alphabets, a small variational distance between them does not imply that the difference between their entropies is small if one of the alphabet sizes is unknown. This fact, seemingly contradictory to the continuity of entropy for finite alphabet, is clarified in the current paper by means of certain bounds on the entropy difference between two probability distributions in terms of the variational distance between them and their alphabet sizes. These bounds are shown to be the tightest possible. The Lagrange multiplier cannot be applied here because the variational distance is not differentiable. We also show how to find the distribution achieving the minimum (or maximum) entropy among those distributions within a given variational distance from any given distribution.

I. Introduction

When we want to find the maximum or minimum value of the Shannon entropy,

$$H(\mathcal{P}) = \sum_{i:p_i>0} p_i \log \frac{1}{p_i},$$

of a probability distribution $\mathcal{P} = \{p_i\}$ subject to some constraints, a typical approach is to apply the Lagrange multiplier [1]. By using differentiation and solving some equations, the solution satisfying the given set of constraints would be obtained. The powerful Lagrange multiplier can usually solve this kind of problems, but sometimes it fails. For example, suppose a probability

S.-W. Ho is with Department of Electrical Engineering, Princeton University, NJ 08544, USA. He was with Department of Information Engineering, The Chinese University of Hong Kong, N.T., Hong Kong when part of this work was done. Email: siuho@princeton.edu

R. W. Yeung is with Department of Information Engineering, The Chinese University of Hong Kong, N.T., Hong Kong. Email: whyeung@ie.cuhk.edu.hk

distribution $\mathcal{P} = \{p_1, p_2, \ldots, p_L\}$ is given and we want to find a probability distribution $\mathcal{Q} = \{q_1, q_2, \ldots, q_L\}$ attaining the maximum entropy subject to the variational distance being less than or equal to ϵ, i.e.

$$V(\mathcal{P}, \mathcal{Q}) = \sum_{i=1}^{L} |p_i - q_i| \leq \epsilon.$$

The Lagrange multiplier cannot be applied here because $V(\mathcal{P}, \mathcal{Q})$ is not differentiable with respect to p_i. Some literatures [2][3] tackled a similar problem for finding an upper bound on the difference of entropies, $|H(\mathcal{P}) - H(\mathcal{Q})|$, subject to $V(\mathcal{P}, \mathcal{Q}) \leq \epsilon$. They used some fundamental inequalities to obtain the bounds but the bounds are not tight. At the same time, we may want to know the minimum entropy and the lower bound on the difference of entropies in the above problems. These mathematical problems will be solved in Section II and their applications in entropy estimation, rate-distortion theory, generalization of the Fano inequality and complexity of random number generation will be shown in Part II of this paper. All the logarithms denoted by $\log$ in this paper are in the same base. The natural logarithm is denoted by $\ln$ and the natural number to the power x is denoted by $\exp(x)$.

II. New Bounds

The following theorem refines Theorem 3 in [4] regarding the discontinuity of the Shannon entropy $H(\cdot)$ with respect to the variational distance. The variational distance between two probability distributions $\mathcal{P} = \{p_1, p_2, \ldots, p_L\}$ and $\mathcal{Q} = \{q_1, q_2, \ldots, q_M\}$ with different support is defined as

$$V(\mathcal{P}, \mathcal{Q}) = \sum_{i=1}^{L} |p_i - q_i| + \sum_{i=L+1}^{M} |q_i|.$$

Theorem 1: Suppose $\delta > 0$ and $\epsilon > 0$ are given. For any probability distribution $\mathcal{P}$ with L probability masses, there exists a sufficient large integer $M \geq L$ and a probability distribution $\mathcal{Q}$ with M probability masses such that the variational distance $V(\mathcal{P}, \mathcal{Q}) < \epsilon$ but $H(\mathcal{Q}) - H(\mathcal{P}) > \delta$.

Proof: Let $\mathcal{P} = \{p_1, p_2, \ldots, p_L\}$ and let

$$\begin{aligned} \mathcal{Q} \;=\; & \left\{ p_1 - \frac{p_1}{\sqrt{\log M}}, p_2 + \frac{p_1}{M\sqrt{\log M}}, \ldots, \right. \\ & \left. p_L + \frac{p_1}{M\sqrt{\log M}}, \frac{p_1}{M\sqrt{\log M}}, \ldots, \frac{p_1}{M\sqrt{\log M}} \right\}. \end{aligned}$$

be a probability distribution with $M+1$ probability masses for $M \geq L$. Then it is readily checked that for any positive ϵ and δ, $V(\mathcal{P}, \mathcal{Q}) < \epsilon$ but $H(\mathcal{Q}) - H(\mathcal{P}) > \delta$ when M is sufficiently large. ■

The four quantities δ, ϵ, L and M play critical roles in Theorem 1 and the relation among them will be explored through a possible application in this section. For $n \in \mathbb{N}$, let

$$\Gamma_n = \left\{ \mathcal{P} = (p_1, p_2, \ldots, p_n) : \sum_{j=1}^{n} p_j = 1, p_j \geq 0, 1 \leq j \leq n \right\},$$

and let

$$\Gamma_\infty = \left\{ \mathcal{P} = (p_1, p_2, \ldots) : \sum_{j=1}^{\infty} p_j = 1, p_j \geq 0, 1 \leq j \right\}.$$

Suppose a probability distribution

$$\mathcal{P} = (p_1, p_2, \ldots, p_L) \in \Gamma_L, \tag{1}$$

where L is finite, is obtained from an iterative algorithm. Let $\mathcal{Q} = (q_1, q_2, \ldots, q_M) \in \Gamma_M$ be the exact solution where $\mathcal{Q}$ and M are unknown. We are interested in the case that $M \geq L$ which can model a truncation error caused by a program implementing the iterative algorithm. Let d_i be real such that

$$\mathcal{Q} = (p_1 + d_1, p_2 + d_2, \ldots, p_L + d_L, d_{L+1}, \ldots, d_M). \tag{2}$$

Then the variational distance between $\mathcal{P}$ and $\mathcal{Q}$ can be written as

$$V(\mathcal{P}, \mathcal{Q}) = \sum_{i=1}^{L} |p_i - (p_i + d_i)| + \sum_{i=L+1}^{M} |d_i| = \sum_{i=1}^{M} |d_i|.$$

Suppose our iterative algorithm has obtained $\mathcal{P}$, within the neighborhood of $\mathcal{Q}$ with respect to the variational distance, say

$$V(\mathcal{P}, \mathcal{Q}) \leq \epsilon. \tag{3}$$

Note that if M is infinite, there exists a $\mathcal{Q}$ such that $H(\mathcal{Q}) - H(\mathcal{P}) = \infty$ from Theorem 3 in [4]. Therefore, no matter how small ϵ is, if the program can only generate probability distributions with finite support to approximate an exact solution with infinite support, then the error in

the estimated entropy can be unbounded. If M is finite but unknown, Theorem 1 says that $H(\mathcal{Q}) - H(\mathcal{P})$ can be any value. However, if M is finite and known, the quantity

$$\sup_{\mathcal{Q}} |H(\mathcal{Q}) - H(\mathcal{P})| \tag{4}$$

is finite and in this paper we will obtain lower and upper bounds on this quantity. For given $\mathcal{P}$, ϵ, L and M, it turns out that the probability distribution $\mathcal{Q}$ achieving the supremum in (4) can be found easily. Note that

$$\begin{aligned} \sum_{i=1}^{M} d_i &= 0 \\ \sum_{i:d_i>0} d_i &= -\sum_{i:d_i<0} d_i, \end{aligned}$$

where d_i's are defined in (2). In order to satisfy $V(\mathcal{P}, \mathcal{Q}) \leq \epsilon$, the choice of d_i must satisfy the condition

$$\sum_{i:d_i>0} d_i = -\sum_{i:d_i<0} d_i \leq \frac{\epsilon}{2}.$$

We also require that $p_i + d_i \geq 0$ for all i. Define the function

$$f(p, \delta) = -(p+\delta)\log(p+\delta) + p \log p,$$

for $p > 0$ and $-p \leq \delta \leq 1 - p$ and define $f(0, \delta) = -\delta \log \delta$. In the proofs of this paper, we will frequently use the following two identities:

1) For $0 \leq \delta' < \delta$ or $0 \geq \delta' > \delta$,

$$f(p, \delta) = f(p, \delta') + f(p + \delta', -\delta' + \delta).$$

2) For $p < p' \leq p + \delta$ or $p > p' \geq p + \delta$,

$$f(p, \delta) = f(p, -p + p') + f(p', -p' + p + \delta).$$

Two other properties of $f(p, \delta)$ are given in the following two lemmas.

Lemma 2: For a fixed $\delta > 0$, $f(p, \delta)$ is a strictly decreasing function on p. In particular,

$$f(0, \delta) = -\delta \log \delta \geq f(p, \delta)$$

for all $0 < p \leq 1 - \delta$.

Lemma 3: For a fixed $\delta < 0$, $f(p, \delta)$ is a strictly increasing function on p. In particular,

$$f(1, \delta) \geq f(p, \delta)$$

for all $-\delta \leq p \leq 1$.

Proof of Lemma 2 and 3

$$\begin{aligned} f'(p,\delta) &= \frac{\partial}{\partial p} f(p,\delta) \\ &= \frac{\partial}{\partial p} (\ln 2)^{-1} (-(p+\delta)\ln(p+\delta) + p \ln p) \\ &= (\ln 2)^{-1} (-1 - \ln(p+\delta) + 1 + \ln p) \\ &= \log \frac{p}{p+\delta}. \end{aligned}$$

For a fixed $\delta > 0$, $f'(p, \delta) < 0$ so that $f(p, \delta)$ is a strictly decreasing function on p and $f(p, \delta)$ is the largest at $p = 0$. For a fixed $\delta < 0$, $f'(p, \delta) > 0$ so that $f(p, \delta)$ is a strictly increasing function on p and $f(p, \delta)$ is the largest at $p = 1$. ■

By Lemma 2 and Lemma 3, we will prove the following two lemmas which will be used frequently in this paper.

Lemma 4: Let $\{p_i, d_i\}$ and $\{p_j^*, d_j^*\}$ be two sets of real numbers such that for all i and j,

$$0 \leq p_i + d_i \leq p_i \leq 1$$

and

$$0 \leq p_j^* + d_j^* \leq p_j^* \leq 1,$$

where d_i and $d_j^* < 0$. If

$$\sum_i d_i = \sum_j d_j^* \tag{5}$$

and

$$\min_i \{p_i + d_i\} \geq \max_j \{p_j^*\}, \tag{6}$$

then

$$\sum_i f(p_i, d_i) \geq \sum_j f(p_j^*, d_j^*).$$

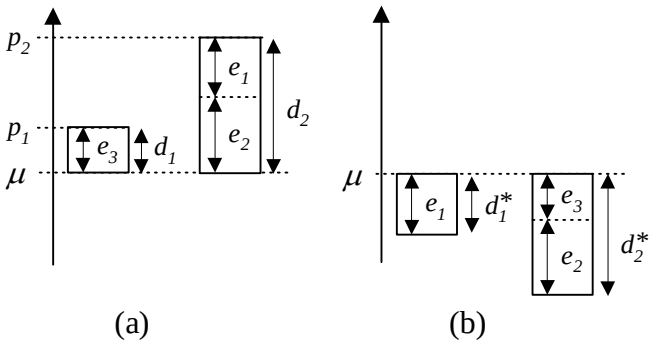

Fig. 1. An illustration of the partition $\{e_i\}$

Proof:

We first consider a simple example in Fig. 1 which illustrates the assumptions in this lemma. The bars in Fig. 1(a) have lengths $\{d_i\}$ while the bars in Fig. 1(b) have lengths $\{d_i^*\}$. Consider a partition $\{e_i\}$, a set of positive real numbers, satisfying that $d_1 = e_3$, $d_2 = e_1 + e_2$, $d_1^* = e_1$ and $d_2^* = e_2 + e_3$. Such partition must exist because $d_1 + d_2 = d_1^* + d_2^*$. Let μ be a real number such that

$$\min_i\{p_i + d_i\} \geq \mu \geq \max_j\{p_j^*\}.$$

Then

$$\begin{aligned} f(p_2, d_2) + f(p_1, d_1) &= f(p_2, -e_1 - e_2) + f(p_1, -e_3) \\ &= f(p_2, -e_1) + f(p_2 - e_1, -e_2) + f(p_1, -e_3) \\ &\geq f(\mu, -e_1) + f(\mu, -e_2) + f(\mu, -e_3), \end{aligned}$$

where the inequality follows from Lemma 3 and $p_i + d_i \geq \mu$ for all i. Together with $\mu \geq p_j^*$ for all j and Lemma 3, we can further show that

$$\begin{aligned} f(p_2, d_2) + f(p_1, d_1) &\geq f(\mu, -e_1) + f(\mu, -e_2) + f(\mu, -e_3) \\ &\geq f(p_1^*, -e_1) + f(p_2^*, -e_3) + f(p_2^* - e_3, -e_2) \\ &= f(p_1^*, -e_1) + f(p_2^*, -e_2 - e_3) \\ &= f(p_1^*, d_1^*) + f(p_2^*, d_2^*). \end{aligned}$$

In general, we can always divide the areas covered by $\{d_i\}$ and $\{d_i^*\}$ into a suitable partition $\{e_k\}$ and show

$$\sum_i f(p_i, d_i) \geq \sum_k f(\mu, -e_k) \geq \sum_j f(p_j^*, d_j^*).$$

■

Lemma 5: Let $\{p_i, d_i\}$ and $\{p_j^*, d_j^*\}$ be two sets of real numbers such that for all i and j,

$$0 \leq p_i \leq p_i + d_i \leq 1$$

and

$$0 \leq p_j^* \leq p_j^* + d_j^* \leq 1,$$

where d_i and $d_j^* > 0$. If

$$\sum_i d_i = \sum_j d_j^* \tag{7}$$

and

$$\min\{p_j^*\} \geq \max\{p_i + d_i\}, \tag{8}$$

then

$$\sum_j f(p_j^*, d_j^*) \leq \sum_i f(p_i, d_i).$$

Proof: Note that

$$-\sum_i f(p_i, d_i) = \sum_i f(p_i + d_i, -d_i).$$

At the same time,

$$\begin{aligned} -\sum_j f(p_j^*, d_j^*) &= \sum_j f(p_j^* + d_j^*, -d_j^*) \\ &\geq \sum_i f(p_i + d_i, -d_i) \\ &= -\sum_i f(p_i, d_i), \end{aligned}$$

where the inequality follows from Lemma 4. Therefore,

$$\sum_j f(p_j^*, d_j^*) \leq \sum_i f(p_i, d_i),$$

and the lemma is proved. ■

By using Lemma 4 and Lemma 5, a process to obtain $\mathcal{Q}$ achieving the supremum in (4) is described in the next two theorems. We will first solve the problem

$$\begin{aligned} \sup_{\mathcal{Q}} \quad & H(\mathcal{Q}) - H(\mathcal{P}) \\ \text{subject to} \quad & V(\mathcal{P}, \mathcal{Q}) \leq \epsilon. \end{aligned}$$

This problem can be solved by any convex optimization method. However, the solution will be shown to be neat and compact in the following theorem. The theorem is useful to prove some theorems in the later part of this paper and in Part II of this paper. In the following, we will use the notations

$$(x)^+ = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0, \end{cases}$$

and

$$(x)^- = \begin{cases} x & \text{if } x < 0 \\ 0 & \text{if } x \geq 0. \end{cases}$$

Theorem 6: Suppose a positive number $\epsilon \leq 2$, $\mathcal{P} = (p_1, p_2, \ldots, p_L) \in \Gamma_L$ and a finite integer $M \geq L$ are given. Let p_i's be sorted in descending order and let $p_{L+1} = p_{L+2} = \cdots = p_M = 0$. Let μ and ν be real such that

$$\sum_{i=1}^{M} (p_i - \mu)^+ = \frac{\epsilon}{2}, \tag{9}$$

and

$$\sum_{i=1}^{M} (\nu - p_i)^+ = \frac{\epsilon}{2}. \tag{10}$$

If $\nu \geq \mu$, let $q_i^* = \frac{1}{M}$ for $1 \leq i \leq M$. Otherwise, let

$$q_i^* = \begin{cases} \mu & \text{if } p_i > \mu \\ p_i & \text{if } \nu \leq p_i \leq \mu \\ \nu & \text{if } p_i < \nu \end{cases}$$

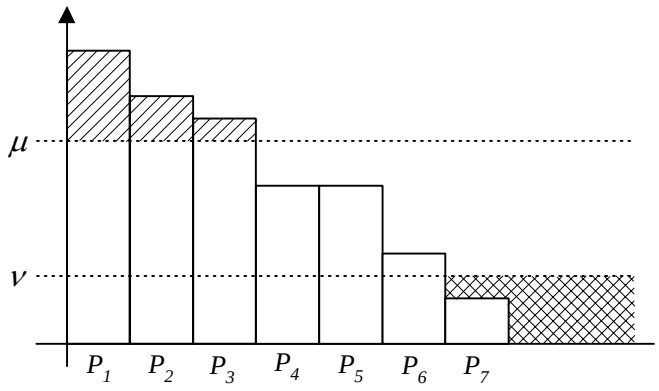

Fig. 2. An example demonstrating the choices of μ and ν according to Theorem 6 where $L = 7$ and $M = 9$. Here, $\mathcal{Q}^* = \{\mu, \mu, \mu, p_4, p_5, p_6, \nu, \nu, \nu\}$.

for $1 \leq i \leq M$ and we denote $\mathcal{Q}^* = \{q_i^*\}$. See Fig. 2 for example. Then for any $\mathcal{Q} \in \Gamma_M$ such that $V(\mathcal{P}, \mathcal{Q}) \leq \epsilon$, we have

$$H(\mathcal{Q}) - H(\mathcal{P}) \leq H(\mathcal{Q}^*) - H(\mathcal{P}).$$

Proof: We follow the definitions of $\mathcal{P}$ and $\mathcal{Q}$ in (1) and (2), respectively. Let d_i^*'s be some real values such that

$$\mathcal{Q}^* = (q_1^*, q_2^*, \ldots, q_M^*) = (p_1 + d_1^*, p_2 + d_2^*, \ldots, p_L + d_L^*, d_{L+1}^*, \ldots, d_M^*),$$

where $\mathcal{Q}^*$ is as specified in the theorem. If $q_i^* = \frac{1}{M}$ for $1 \leq i \leq M$, then $\nu \geq \mu$. For any $\mathcal{Q} \in \Gamma_M$,

$$H(\mathcal{Q}) - H(\mathcal{P}) \leq \log M - H(\mathcal{P}) = H(\mathcal{Q}^*) - H(\mathcal{P}),$$

and

$$V(\mathcal{P}, \mathcal{Q}^*) \leq \epsilon.$$

Otherwise, we know that

$$\sum_{i:d_i^*>0} d_i^* = -\sum_{i:d_i^*<0} d_i^* = \frac{\epsilon}{2},$$

and $\sum_{i=1}^{M} |d_i^*| = \epsilon$. We first consider $V(\mathcal{P}, \mathcal{Q}) = \epsilon$. Then

$$\sum_{i:d_i^*<0} d_i^* = \sum_{i:d_i<0} d_i = -\frac{\epsilon}{2},$$

and

$$\sum_{i:d_i^*>0} d_i^* = \sum_{i:d_i>0} d_i = \frac{\epsilon}{2}.$$

Note that

$$\begin{aligned}
&(H(\mathcal{Q}) - H(\mathcal{P})) - (H(\mathcal{Q}^*) - H(\mathcal{P})) \\
&= \left(\sum_i f(p_i, d_i)\right) - \left(\sum_i f(p_i, d_i^*)\right) \\
&= \left(\sum_{i:d_i<0} f(p_i, d_i) - \sum_{i:d_i^*<0} f(p_i, d_i^*)\right) + \left(\sum_{i:d_i>0} f(p_i, d_i) - \sum_{i:d_i^*>0} f(p_i, d_i^*)\right). \qquad (11)
\end{aligned}$$

Consider the first bracket on the R.H.S. of (11),

$$\begin{aligned}
&\sum_{i:d_i<0} f(p_i, d_i) - \sum_{i:d_i^*<0} f(p_i, d_i^*) \\
&= \sum_i f(p_i, (d_i)^-) - \sum_i f(p_i, (d_i^*)^-) \\
&= \sum_i f(p_i + (d_i^*)^-, (d_i)^- - (d_i^*)^-) \\
&= \sum_{i:(d_i)^- - (d_i^*)^- > 0} f(p_i + (d_i^*)^-, (d_i)^- - (d_i^*)^-) \\
&\quad + \sum_{i:(d_i)^- - (d_i^*)^- < 0} f(p_i + (d_i^*)^-, (d_i)^- - (d_i^*)^-) \\
&= -\sum_{i:(d_i)^- - (d_i^*)^- > 0} f(p_i + (d_i)^-, (d_i^*)^- - (d_i)^-) \\
&\quad + \sum_{i:(d_i)^- - (d_i^*)^- < 0} f(p_i + (d_i^*)^-, (d_i)^- - (d_i^*)^-).
\end{aligned}$$

By the definition of q_i and the assumption that $\mu > \nu$, we have

$$p_i + (d_i^*)^- \le p_i + d_i^* = q_i^* \le \mu$$

for all i. Then by Lemma 3,

$$\begin{aligned}
&\sum_{i:d_i<0} f(p_i, d_i) - \sum_{i:d_i^*<0} f(p_i, d_i^*) \\
&\le -\sum_{i:(d_i)^- - (d_i^*)^- > 0} f(p_i + (d_i)^-, (d_i^*)^- - (d_i)^-) \\
&\quad + \sum_{i:(d_i)^- - (d_i^*)^- < 0} f(\mu, (d_i)^- - (d_i^*)^-). \qquad (12)
\end{aligned}$$

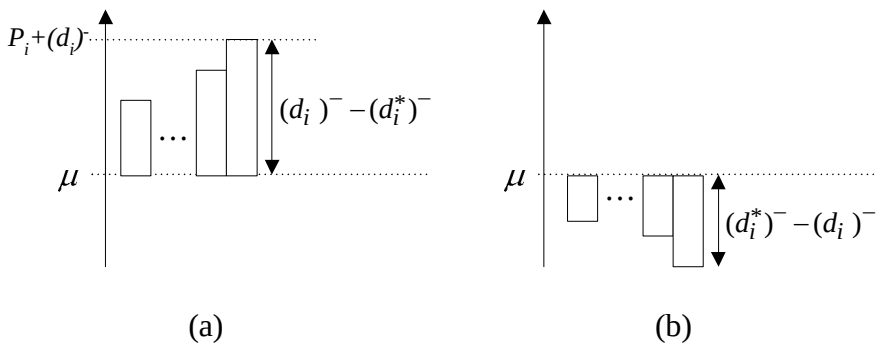

Fig. 3. A picture illustrating the magnitudes of the variables in (12)

In the first summation of (12), the summation is over all i satisfying

$$(d_i^*)^- < (d_i)^- \leq 0. \tag{13}$$

Since $(d_i^*)^- < 0$, we have $d_i^* < 0$ and $p_i + (d_i^*)^- = \mu$. At the same time,

$$p_i + (d_i)^- > p_i + (d_i^*)^- = \mu. \tag{14}$$

By using the relations in (13) and (14), the terms inside $f(\cdot,\cdot)$ in the first summation and the second summation of (12) are picturised in Fig. 3(a) and Fig. 3(b), respectively. Note that the area covered by the bar chart in Fig. 3(a) and Fig. 3(b) are the same because

$$\begin{aligned} & \sum_{i:(d_i)^- - (d_i^*)^- > 0} ((d_i)^- - (d_i^*)^-) - \sum_{i:(d_i)^- - (d_i^*)^- < 0} ((d_i^*)^- - (d_i)^-) \\ = \; & \sum_i (d_i)^- - \sum_i (d_i^*)^- \\ = \; & 0. \end{aligned}$$

By Lemma 4, it is readily seen that

$$\sum_{i:(d_i)^- - (d_i^*)^- > 0} f(p_i + (d_i)^-, (d_i^*)^- - (d_i)^-) \geq \sum_{i:(d_i)^- - (d_i^*)^- < 0} f(\mu, (d_i)^- - (d_i^*)^-).$$

By (12), we have

$$\sum_{i:d_i<0} f(p_i, d_i) \leq \sum_{i:d_i^*<0} f(p_i, d_i^*). \tag{15}$$

Consider the second bracket on the R.H.S. of (11),

$$\begin{aligned}
&\sum_{i:d_i>0} f(p_i,d_i) - \sum_{i:d_i^*>0} f(p_i,d_i^*) \\
&= \sum_{i:d_i>0} f(p_i,(d_i)^+) - \sum_{i:d_i^*>0} f(p_i,(d_i^*)^+) \\
&= \sum_{i} f(p_i,(d_i)^+) - \sum_{i} f(p_i,(d_i^*)^+) \\
&= \sum_{i} f(p_i+(d_i^*)^+,(d_i)^+ - (d_i^*)^+) \\
&= \sum_{i:(d_i)^+-(d_i^*)^+>0} f(p_i+(d_i^*)^+,(d_i)^+ - (d_i^*)^+) + \\
&\quad \sum_{i:(d_i)^+-(d_i^*)^+<0} f(p_i+(d_i^*)^+,(d_i)^+ - (d_i^*)^+) \\
&= \sum_{i:(d_i)^+-(d_i^*)^+>0} f(p_i+(d_i^*)^+,(d_i)^+ - (d_i^*)^+) - \\
&\quad \sum_{i:(d_i)^+-(d_i^*)^+<0} f(p_i+(d_i)^+,(d_i^*)^+ - (d_i)^+)
\end{aligned}$$

Since $p_i + (d_i^*)^+ \geq p_i + d_i^* = q_i \geq \nu$, by Lemma 2, we have

$$\begin{aligned}
&\sum_{i:d_i>0} f(p_i,d_i) - \sum_{i:d_i^*>0} f(p_i,d_i^*) \\
&\leq \sum_{i:(d_i)^+-(d_i^*)^+>0} f(\nu,(d_i)^+ - (d_i^*)^+) - \sum_{i:(d_i)^+-(d_i^*)^+<0} f(p_i+(d_i)^+,(d_i^*)^+ - (d_i)^+) \quad (16)
\end{aligned}$$

In the second summation of (16), we have

$$(d_i^*)^+ > (d_i)^+ \geq 0. \quad (17)$$

Since $(d_i^*)^+ > 0$, we have $d_i^* > 0$ and $p_i + (d_i^*)^+ = \nu$. At the same time,

$$p_i + (d_i)^+ < p_i + (d_i^*)^+ = \nu. \quad (18)$$

By using the relations in (17) and (18), the terms inside $f(\cdot,\cdot)$ in the first summation and the second summation of (16) are picturised in Fig. 4 (a) and Fig. 4 (b), respectively. Note that the area covered by the bar chart in Fig. 4 (a) and Fig. 4 (b) are the same because

$$\begin{aligned}
&\sum_{i:(d_i)^+-(d_i^*)^+>0} ((d_i)^+ - (d_i^*)^+) - \sum_{i:(d_i)^+-(d_i^*)^+<0} ((d_i^*)^+ - (d_i)^+) \\
&= \sum_i (d_i)^+ - \sum_i (d_i^*)^+ \\
&= 0.
\end{aligned}$$

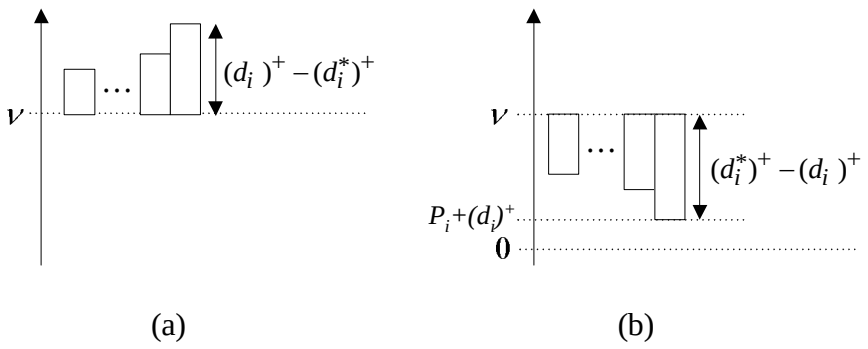

Fig. 4. A picture illustrating the magnitudes of the variables in (16)

By Lemma 5, it is readily seen that

$$\sum_{i:(d_i)^+-(d_i^*)^+<0} f(p_i+(d_i)^+,(d_i^*)^+-(d_i)^+) \geq \sum_{i:(d_i)^+-(d_i^*)^+>0} f(\nu,(d_i)^+-(d_i^*)^+).$$

By (16), we have

$$\sum_{i:d_i>0} f(p_i,d_i) \leq \sum_{i:d_i^*>0} f(p_i,d_i^*). \tag{19}$$

By putting (15) and (19) into (11), for any $\mathcal{Q} \in \Gamma_M$,

$$\begin{aligned} H(\mathcal{Q})-H(\mathcal{P}) &= \sum_i f(p_i,d_i) \\ &\leq \sum_i f(p_i,d_i^*) \\ &= H(\mathcal{Q}^*)-H(\mathcal{P}), \end{aligned}$$

which proves the theorem for $V(\mathcal{P},\mathcal{Q})=\epsilon$.

If $V(\mathcal{P},\mathcal{Q})=\eta<\epsilon$, then we first find $\mathcal{Q}^{**}$ and $\mathcal{Q}^*$ according to the definition in this theorem with respect to η and ϵ, respectively. The previous part of this proof has already shown that

$$H(\mathcal{Q})-H(\mathcal{P}) \leq H(\mathcal{Q}^{**})-H(\mathcal{P}).$$

The proof is completed if we can show that $H(\mathcal{Q}^*) \geq H(\mathcal{Q}^{**})$. Note that for $\delta > 0$, $p > \frac{1}{M} + \delta$ and $p' < \frac{1}{M} - \delta$,

$$\begin{aligned}
& f(p, -\delta) + f(p', \delta) && (20)\\
&\geq f\left(\frac{1}{M} + \delta, -\delta\right) + f(p', \delta) && (21)\\
&\geq f\left(\frac{1}{M} + \delta, -\delta\right) + f\left(\frac{1}{M} - \delta, \delta\right) && (22)\\
&= -\frac{2}{M}\log\frac{1}{M} + \left(\frac{1}{M} + \delta\right)\log\left(\frac{1}{M} + \delta\right) + \left(\frac{1}{M} - \delta\right)\log\left(\frac{1}{M} - \delta\right)\\
&\geq -\frac{2}{M}\log\frac{1}{M} + \frac{2}{M}\log\frac{1}{M} && (23)\\
&= 0,
\end{aligned}$$

where (21) follows from Lemma 3, (22) follows from Lemma 2 and (23) follows from that fact that $x \log x$ is a strictly convex function. Therefore, the additional decrements of the large probability masses and increments of the small probability masses in $\mathcal{Q}^{**}$ with respect to $\mathcal{Q}^*$ makes $H(\mathcal{Q}^*) \geq H(\mathcal{Q}^{**})$. Finally,

$$H(\mathcal{Q}) - H(\mathcal{P}) \leq H(\mathcal{Q}^{**}) - H(\mathcal{P}) \leq H(\mathcal{Q}^*) - H(\mathcal{P}),$$

and the proof is completed. ■

From the proof of Theorem 6, it is readily checked the following corollary.

Corollary 7: For any probability distribution $\mathcal{P}$, let $\mathcal{Q}$ and $\mathcal{Q}'$ be the $\mathcal{Q}^*$ as specified in Theorem 6 with $V(\mathcal{P}, \mathcal{Q}) \leq \epsilon$ and $V(\mathcal{P}, \mathcal{Q}') \leq \epsilon'$, respectively. Assume $\epsilon' < \epsilon$. If $H(\mathcal{Q}) < \log M$, then

$$V(\mathcal{P}, \mathcal{Q}) = \epsilon,$$

$$V(\mathcal{P}, \mathcal{Q}') = \epsilon',$$

and

$$H(\mathcal{Q}) > H(\mathcal{Q}').$$

In Theorem 6, when we find the distribution $\mathcal{Q}$ that maximizes $H(\mathcal{Q}) - H(\mathcal{P})$ subject to $V(\mathcal{P}, \mathcal{Q}) \leq \epsilon$, it is necessary to impose an upper bound on the alphabet size of $\mathcal{Q}$, because otherwise $H(\mathcal{Q}) - H(\mathcal{P})$ is unbounded. However, when we find the distribution $\mathcal{Q}$ that maximizes

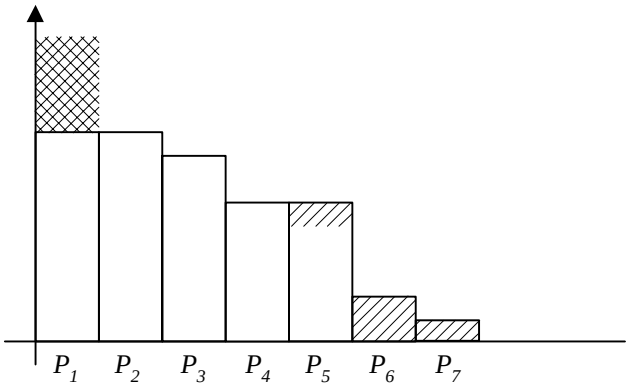

Fig. 5. An example demonstrating the choices of q_i^*'s according to Theorem 8 where $L = 7$. Here, $\mathcal{Q}^* = \{p_1 + \frac{\epsilon}{2}, p_2, p_3, p_4, p_5 + p_6 + p_7 - \frac{\epsilon}{2}\}$.

$H(\mathcal{P}) - H(\mathcal{Q})$ subject to $V(\mathcal{P}, \mathcal{Q}) \leq \epsilon$, it is not necessary to impose an upper bound on the alphabet size of $\mathcal{Q}$, as we will see in the next theorem. In Part II of this paper, this property is used to give a simple proof that entropy is lower semi-continuous. In the following, we will use majorization [5] to give a simple proof. We say that $\mathcal{Q}^* = \{q_i^*\} \in \Gamma_M$ majorizes $\mathcal{Q} = \{q_i\} \in \Gamma_M$, where q_i^*'s and q_i's are sorted in descending order, if

$$\sum_{i=1}^{n} q_i^* \geq \sum_{i=1}^{n} q_i$$

for $1 \leq n \leq M$. Moreover, a function $g(\cdot)$ is strictly Suhur-concave if $g(\mathcal{Q}) > g(\mathcal{Q}^*)$ whenever $\mathcal{Q}^*$ majorizes $\mathcal{Q}$.

Theorem 8: Suppose a positive number $\epsilon \leq 2$ and $\mathcal{P} = (p_1, p_2, \ldots, p_L) \in \Gamma_L$ are given where L can be infinity. Assume p_i's are sorted in descending order. If $1 - p_1 \leq \frac{\epsilon}{2}$, let $\mathcal{Q}^* = (1, 0, \ldots, 0) \in \Gamma_L$. Otherwise, let K be the largest integer such that

$$\sum_{i=K}^{L} p_i \geq \frac{\epsilon}{2}.$$

Let

$$q_i^* = \begin{cases} p_1 + \frac{\epsilon}{2} & \text{if } i = 1 \\ p_i & \text{if } 1 < i < K \\ \sum_{i=K}^{L} p_i - \frac{\epsilon}{2} & \text{if } i = K. \end{cases}$$

We denote $\mathcal{Q}^* = \{q_i^*\}$. See Fig. 5 for example. For any $\mathcal{Q}$ such that $V(\mathcal{P}, \mathcal{Q}) \leq \epsilon$, we have

$$H(\mathcal{P}) - H(\mathcal{Q}) \leq H(\mathcal{P}) - H(\mathcal{Q}^*).$$

Proof: We follow the definitions of $\mathcal{P}$ and $\mathcal{Q}^*$ as specified in the theorem. Consider $\mathcal{Q} \in \Gamma_M$ with $V(\mathcal{P}, \mathcal{Q}) \leq \epsilon$ where $M \geq L$ is a positive integer or $M = \infty$. If $1 - p_1 \leq \frac{\epsilon}{2}$, $\mathcal{Q}^* = (1, 0, \ldots, 0) \in \Gamma_L$. Then

$$H(\mathcal{P}) - H(\mathcal{Q}) \leq H(\mathcal{P}) - 0 = H(\mathcal{P}) - H(\mathcal{Q}^*)$$

and

$$V(\mathcal{P}, \mathcal{Q}^*) \leq \epsilon.$$

Otherwise, we know that

$$V(\mathcal{P}, \mathcal{Q}^*) = \epsilon.$$

and

$$d_1^* = \frac{\epsilon}{2}.$$

Since $V(\mathcal{P}, \mathcal{Q}) \leq \epsilon$, we have

$$\sum_{i:d_i>0} d_i = -\sum_{i:d_i<0} d_i \leq \frac{\epsilon}{2},$$

and

$$\sum_{i=1}^{n} d_i \leq \frac{\epsilon}{2} \tag{24}$$

for $1 \leq n \leq M$. Let $\tilde{\mathcal{Q}} = \{\tilde{q_1}, \tilde{q_2}, \ldots, \tilde{q_M}\} \in \Gamma_M$ contain all the probability masses in $\mathcal{Q}$ (c.f. (2)) but be sorted in descending order. At the same time, add $q_{K+1}^* = q_{K+2}^* = \cdots = q_M^* = 0$ into $\mathcal{Q}^*$ such that $\mathcal{Q}^* \in \Gamma_M$. Together with (24), it is readily seen that for any $n < K$,

$$\sum_{i=1}^{n} q_i^* = \sum_{i=1}^{n} p_i + \frac{\epsilon}{2} \geq \sum_{i=1}^{n} (p_i + d_i) \geq \sum_{i=1}^{n} \tilde{q}_i.$$

Furthermore, for $K \leq n \leq M$,

$$\sum_{i=1}^{n} q_i^* = 1 \geq \sum_{i=1}^{n} \tilde{q}_i.$$

Therefore, $\mathcal{Q}^*$ majorizes $\mathcal{Q}$. Since entropy $H(\cdot)$ is strictly Suhur-concave [5], we have

$$H(\mathcal{Q}) \geq H(\mathcal{Q}^*),$$

which means

$$H(\mathcal{P}) - H(\mathcal{Q}) \leq H(\mathcal{P}) - H(\mathcal{Q}^*).$$

■

The nice property that only p_1 is increased in Theorem 8 helps to show that the $\mathcal{Q}^*$ in Theorem 8 majorizes all the feasible $\mathcal{Q}$. Unfortunately, there is no such nice property for the $\mathcal{Q}^*$ in Theorem 6 and therefore majorization may not shorten the proof of Theorem 6. From Theorem 8, it is readily checked the following corollary.

Corollary 9: For any probability distribution $\mathcal{P}$, let $\mathcal{Q}$ and $\mathcal{Q}'$ be the $\mathcal{Q}^*$ as specified in Theorem 8 with $V(\mathcal{P}, \mathcal{Q}) \leq \epsilon$ and $V(\mathcal{P}, \mathcal{Q}') \leq \epsilon'$, respectively. Assume $\epsilon' < \epsilon$. If $1 - p_1 > \frac{\epsilon}{2}$, then

$$V(\mathcal{P}, \mathcal{Q}) = \epsilon,$$

$$V(\mathcal{P}, \mathcal{Q}') = \epsilon',$$

and

$$H(\mathcal{Q}') > H(\mathcal{Q}).$$

Now, we are readily to obtain an upper bound on (4). Note that for any given $\mathcal{P}$,

$$\begin{aligned}
&\sup_{\mathcal{Q}} |H(\mathcal{Q}) - H(\mathcal{P})| \\
&= \max\left\{\sup_{\mathcal{Q}} \left(H(\mathcal{Q}) - H(\mathcal{P})\right), -\inf_{\mathcal{Q}} \left(H(\mathcal{Q}) - H(\mathcal{P})\right)\right\} \\
&= \max\left\{\sup_{\mathcal{Q}} \left(H(\mathcal{Q}) - H(\mathcal{P})\right), \sup_{\mathcal{Q}} \left(H(\mathcal{P}) - H(\mathcal{Q})\right)\right\} \\
&= \max\{H(\mathcal{Q}^+) - H(\mathcal{P}), H(\mathcal{P}) - H(\mathcal{Q}^-)\},
\end{aligned}$$

where

$$\mathcal{Q}^+ = \arg\max_{\mathcal{Q}} \left(H(\mathcal{Q}) - H(\mathcal{P})\right)$$

and

$$\mathcal{Q}^- = \arg\max_{\mathcal{Q}} \left(H(\mathcal{P}) - H(\mathcal{Q})\right)$$

can be obtained from Theorem 6 and Theorem 8, respectively. Then the upper bound on (4) is obtained by comparing $H(\mathcal{Q}^+) - H(\mathcal{P})$ with $H(\mathcal{P}) - H(\mathcal{Q}^-)$. Note that if $\mathcal{P}$ is an empirical distribution obtained from a source, Theorem 6 and Theorem 8 can give us an estimated range of the true entropy. This will give us the confidence interval of the true entropy in Part II of this paper.

The value of (4) can be obtained by Theorem 6 and Theorem 8 and the value depends not only on M but also on $\mathcal{P}$. It is also interesting to obtain an upper bound on (4), which is independent of $\mathcal{P}$. Such a bound can be used to determine a stopping condition in an iterative algorithm for obtaining $\mathcal{P}$ as an approximation of an unknown distribution $\mathcal{Q}$ whose alphabet size M is known. We will first prove a more general result by assuming the probability masses in $\mathcal{P}$ and $\mathcal{Q}$ are less than a. In order to make the expression simple, we assumed $a = \frac{1}{N}$ for an integer N.

Theorem 10: If $\mathcal{P} \in \Gamma_L$ and $\mathcal{Q} \in \Gamma_M$ with $M \geq L$ being two probability distributions such that $V(\mathcal{P}, \mathcal{Q}) \leq \epsilon$ and the probability masses in $\mathcal{P}$ and $\mathcal{Q}$ are less than $a = \frac{1}{N}$ for an integer $N \leq L$, then

$$|H(\mathcal{Q}) - H(\mathcal{P})| \leq \begin{cases} H\left(\{\frac{\epsilon}{2}, 1-\frac{\epsilon}{2}\}\right) + \frac{\epsilon}{2}\log\frac{M-N}{N} & 0 < \frac{\epsilon}{2} < \frac{M-N}{M} \\ \log M - \log N & \frac{\epsilon}{2} \geq \frac{M-N}{M}. \end{cases}$$

Proof: We follow the definitions of $\mathcal{P}$ and $\mathcal{Q}$ in (1) and (2), respectively. Assume p_i's are sorted in descending order. Let $\gamma = \{a, a, \ldots, a, 0, 0, \ldots, 0\} \in \Gamma_L$ where the first N probability masses are equal to a. We first prove that $H(\mathcal{P}) \geq H(\gamma)$. Consider

$$\begin{aligned} H(\mathcal{P}) - H(\gamma) &= \sum_{i=1}^{L} f(0, p_i) - \sum_{i=1}^{N} f(0, a) \\ &= \sum_{i=N+1}^{L} f(0, p_i) - \sum_{i=1}^{N} f(p_i, -p_i + a). \end{aligned} \tag{25}$$

Before Lemma 5 can be applied, we need to check two conditions. It is easily checked that

$$\sum_{i=N+1}^{L} p_i = 1 - \sum_{i=1}^{N} p_i = \sum_{i=1}^{N}(-p_i + a),$$

and

$$\max_{i:N+1\le i\le L}\{p_i\} \le p_N = \min_{i:1\le i\le N}\{p_i\}.$$

By Lemma 5 and (25), we have

$$H(\mathcal{P}) \ge H(\gamma).$$

Together with $H(\mathcal{Q}) \le \log M$, the upper bound

$$|H(\mathcal{Q}) - H(\mathcal{P})| \le \log M - H(\gamma) = \log M - \log N$$

is always valid. Since

$$V\left(\gamma, \left\{\frac{1}{M}, \frac{1}{M}, \ldots, \frac{1}{M}\right\}\right) = 2\cdot\frac{M-N}{M},$$

we now consider $0 < \frac{\epsilon}{2} < \frac{M-N}{M}$ for a tighter bound on $|H(\mathcal{P}) - H(\mathcal{Q})|$. For those $\mathcal{Q}'$ satisfying $V(\gamma, \mathcal{Q}') \le \epsilon$,

$$\begin{aligned}
&\max_{\mathcal{Q}'}(H(\mathcal{Q}') - H(\gamma)) \\
&= H\left(\left\{a - \frac{a\epsilon}{2}, \ldots, a - \frac{a\epsilon}{2}, \frac{\epsilon}{2(M-N)}, \ldots, \frac{\epsilon}{2(M-N)}\right\}\right) - H(\gamma) \qquad (26)\\
&= Nf\left(a, -\frac{a\epsilon}{2}\right) + (M-N)f\left(0, \frac{\epsilon}{2(M-N)}\right), \qquad (27)
\end{aligned}$$

where (26) follows from Theorem 6. On the other hand, for any $\mathcal{P} \in \Gamma_L$ and $\mathcal{Q}'' = \{q_i''\} \in \Gamma_M$ with $M \ge L$ such that $V(\mathcal{P}, \mathcal{Q}'') \le \epsilon$, we have

$$H(\mathcal{Q}'') - H(\mathcal{P}) \le H(\mathcal{Q}) - H(\mathcal{P}),$$

where $\mathcal{Q}$ is the $\mathcal{Q}^*$ as specified in Theorem 6 subject to $V(\mathcal{P}, \mathcal{Q}) \le \epsilon$. We first consider $V(\mathcal{P}, \mathcal{Q}) = \epsilon$ and we will show

$$H(\mathcal{Q}) - H(\mathcal{P}) \le Nf\left(a, -\frac{a\epsilon}{2}\right) + (M-N)f\left(0, \frac{\epsilon}{2(M-N)}\right).$$

When $\mathcal{Q}$ is obtained from Theorem 6, μ and ν are found. In Fig.2, it is easy to see that μ obtained in case i) $a = p_1 = p_2 = p_3$ is larger than μ obtained in case ii) $a = p_1 > p_2 > p_3$. Let x be the μ obtained in case i). We have $(a-x)N = \frac{\epsilon}{2}$, so that $x = a - \frac{a\epsilon}{2}$. Hence, for all i,

$$p_i + d_i \le \mu \le x = a - \frac{a\epsilon}{2}. \qquad (28)$$

Moreover, the ν obtained in case i) $p_7 = 0$ is smaller than the ν obtained in case ii) $p_7 > 0$. Let y be the ν obtained in case i). Since $\mathcal{P}$ has al least N positive probability masses, we have $y(M-N) = \frac{\epsilon}{2}$, so that $y = \frac{\epsilon}{M-N}$. Hence, for all i,

$$p_i + d_i \geq \gamma \geq y = \frac{\epsilon}{2(M-N)}. \tag{29}$$

Note that

$$\begin{aligned} & H(\mathcal{Q}) - H(\mathcal{P}) - Nf\left(a, -\frac{a\epsilon}{2}\right) - (M-N)f\left(0, \frac{\epsilon}{2(M-N)}\right) \\ = & \sum_i f(p_i, d_i) - Nf\left(a, -\frac{a\epsilon}{2}\right) - (M-N)f\left(0, \frac{\epsilon}{2(M-N)}\right) \\ = & \left(\sum_{i:d_i<0} f(p_i, d_i) - Nf\left(a, -\frac{a\epsilon}{2}\right)\right) \\ & + \left(\sum_{i:d_i>0} f(p_i, d_i) - (M-N)f\left(0, \frac{\epsilon}{2(M-N)}\right)\right). \end{aligned} \tag{30}$$

We first consider the first bracket on the R.H.S. of (30). Define two sets of positive integers

$$S_0 = \left\{i \leq N : d_i < 0 \text{ and } p_i \geq a - \frac{a\epsilon}{2}\right\}$$

and

$$S_1 = \{i : d_i < 0 \text{ and } i \notin S_0\}.$$

Denote the sizes of the sets S_0 and S_1 by $|S_0|$ and $|S_1|$, respectively. Then

$$\begin{aligned}
&\sum_{i:d_i<0} f(p_i,d_i) - Nf\left(a,-\frac{a\epsilon}{2}\right) \\
&= \sum_{i\in S_0}\left(f(p_i,d_i) - f\left(a,-\frac{a\epsilon}{2}\right)\right) + \sum_{i\in S_1} f(p_i,d_i) - (N-|S_0|)f\left(a,-\frac{a\epsilon}{2}\right) \\
&= \sum_{i\in S_0}\left(f\left(p_i,-p_i+a-\frac{a\epsilon}{2}\right) + f\left(a-\frac{a\epsilon}{2},-a+\frac{a\epsilon}{2}+d_i+p_i\right)\right. \\
&\quad \left. -f(a,-a+p_i) - f\left(p_i,-p_i+a-\frac{a\epsilon}{2}\right)\right) + \sum_{i\in S_1} f(p_i,d_i) - (N-|S_0|)f\left(a,-\frac{a\epsilon}{2}\right) \\
&= \sum_{i\in S_0}\left(f\left(a-\frac{a\epsilon}{2},-a+\frac{a\epsilon}{2}+d_i+p_i\right) - f(a,p_i-a)\right) + \sum_{i\in S_1} f(p_i,d_i) - \\
&\quad (N-|S_0|)f\left(a,-\frac{a\epsilon}{2}\right) \\
&= \left(\sum_{i\in S_0} f\left(a-\frac{a\epsilon}{2},-a+\frac{a\epsilon}{2}+d_i+p_i\right) + \sum_{i\in S_1} f(p_i,d_i)\right) \\
&\quad - \left(\sum_{i\in S_0} f(a,p_i-a) - (N-|S_0|)f\left(a,-\frac{a\epsilon}{2}\right)\right), \qquad (31)
\end{aligned}$$

where $-a+\frac{a\epsilon}{2}+d_i+p_i<0$ follows from (28). Before we can apply Lemma 4 to show the R.H.S. of (31) less than 0, we need to check the conditions in (5) and (6). Note that the canceled terms do not affect that

$$\sum_{i\in S_0}\left(-a+\frac{a\epsilon}{2}+d_i+p_i\right) + \sum_{i\in S_1} d_i = \sum_{i\in S_0}(p_i-a) + (N-|S_0|)\left(-\frac{a\epsilon}{2}\right).$$

Therefore, (5) is satisfied. Once (6) is also checked, the fact that the R.H.S. of (31) is less than 0 can be seen immediately from Lemma 4. The checking of (6) is done in the following three cases.

Case 1: $|S_0|=N$. The R.H.S. of (31) becomes

$$\left(\sum_{i\in S_0} f\left(a-\frac{a\epsilon}{2},-a+\frac{a\epsilon}{2}+d_i+p_i\right) + \sum_{i\in S_1} f(p_i,d_i)\right) - \sum_{i\in S_0} f(a,p_i-a).$$

We have

$$\begin{aligned}
\max\left\{a-\frac{a\epsilon}{2}, \max_{i\in S_1} p_i\right\} &\le \max\left\{a-\frac{a\epsilon}{2}, p_N\right\} \\
&= p_N \qquad (32) \\
&\le \min_{i\in S_0} p_i,
\end{aligned}$$

where (32) follows from $p_N \in S_0$.

Case 2: $|S_0| < N$ and $|S_1| = 0$. The R.H.S. of (31) becomes

$$\left(\sum_{i\in S_0} f\left(a-\frac{a\epsilon}{2}, -a+\frac{a\epsilon}{2}+d_i+p_i\right)\right) - \left(\sum_{i\in S_0} f(a, p_i-a) - (N-|S_0|)f\left(a, -\frac{a\epsilon}{2}\right)\right).$$

Then

$$\min\left\{\min_{i\in S_0}\{a+p_i-a\}, a-\frac{a\epsilon}{2}\right\} \geq \min\left\{a-\frac{a\epsilon}{2}, a-\frac{a\epsilon}{2}\right\} = a-\frac{a\epsilon}{2}.$$

Case 3: $|S_0| < N$ and $|S_1| > 0$. Let $b = |S_0|$. Since p_i's are sorted in descending order,

$$\max_{i\in S_1} p_i = p_{b+1}.$$

Since $p_{b+1} \notin S_0$ but $b+1 \leq N$, it must be

$$p_{b+1} < a - \frac{a\epsilon}{2}.$$

Case 3a: $N > |S_0| = b = 0$ and $|S_1| > 0$. Then

$$a - \frac{a\epsilon}{2} \geq p_1 \geq \max_{i\in S_1} p_i.$$

Case 3b: $N > |S_0| = b > 0$ and $|S_1| > 0$. Then

$$\begin{aligned}\max\left\{a-\frac{a\epsilon}{2}, \max_{i\in S_1} p_i\right\} &= \max\left\{a-\frac{a\epsilon}{2}, p_{b+1}\right\} \\ &\leq \max\left\{a-\frac{a\epsilon}{2}, a-\frac{a\epsilon}{2}\right\} \\ &= a-\frac{a\epsilon}{2}\end{aligned}$$

and

$$\begin{aligned}\min\left\{\min_{i\in S_0} p_i, a-\frac{a\epsilon}{2}\right\} &\geq \min\left\{a-\frac{a\epsilon}{2}, a-\frac{a\epsilon}{2}\right\} \\ &= a-\frac{a\epsilon}{2} \\ &\geq \max\left\{a-\frac{a\epsilon}{2}, \max_{i\in S_1} p_i\right\}.\end{aligned}$$

Therefore, (6) is checked in all possible cases and we can apply Lemma 4 to (31) and show that

$$\sum_{i:d_i<0} f(p_i, d_i) - Nf\left(a, -\frac{a\epsilon}{2}\right) \leq 0. \tag{33}$$

Now, we consider the second bracket in (30). Define

$$S_2 = \left\{i > N : d_i > 0 \text{ and } p_i \leq \frac{\epsilon}{2(M-N)}\right\}$$

and

$$S_3 \;=\; \{i : d_i > 0 \text{ and } i \notin S_2\}\,.$$

Then

$$\begin{aligned}
&\sum_{i:d_i>0} f(p_i,d_i) - (M-N)f\left(0,\frac{\epsilon}{2(M-N)}\right)\\
&= \sum_{i\in S_2}\left(f\left(p_i,-p_i+\frac{\epsilon}{2(M-N)}\right)+f\left(\frac{\epsilon}{2(M-N)},-\frac{\epsilon}{2(M-N)}+p_i+d_i\right)\right)\\
&\quad+\sum_{i\in S_3} f(p_i,d_i)-\sum_{i\in S_2} f\left(0,\frac{\epsilon}{2(M-N)}\right)-(M-N-|S_2|)f\left(0,\frac{\epsilon}{2(M-N)}\right)\\
&= \left(\sum_{i\in S_2} f\left(\frac{\epsilon}{2(M-N)},-\frac{\epsilon}{2(M-N)}+p_i+d_i\right)+\sum_{i\in S_3} f(p_i,d_i)\right)\\
&\quad-\left(\sum_{i\in S_2} f\left(0,p_i\right)+(M-N-|S_2|)f\left(0,\frac{\epsilon}{2(M-N)}\right)\right), \qquad (34)
\end{aligned}$$

where $-\frac{\epsilon}{2(M-N)}+p_i+d_i>0$ follows from (29). Before we can apply Lemma 5 to show the R.H.S. of (34) less than 0, we need to check the conditions in (7) and (8). Note that the canceled terms will not affect

$$\sum_{i\in S_2}\left(-\frac{\epsilon}{2(M-N)}+p_i+d_i\right)+\sum_{i\in S_3} d_i=\sum_{i\in S_2} p_i+(M-N-|S_2|)\cdot\frac{\epsilon}{2(M-N)}.$$

Therefore, (7) is satisfied. Once (8) is also checked, the fact that the R.H.S. of (34) is less than 0 can be seen immediately from Lemma 5. The checking of (8) is done in the following three cases.

Case 1: $|S_2|=0$. The R.H.S. of (34) becomes

$$\sum_{i\in S_3} f(p_i,d_i)-(M-N)f\left(0,\frac{\epsilon}{2(M-N)}\right).$$

Then

$$\min_{i\in S_3} p_i \geq \frac{\epsilon}{2(M-N)}.$$

Case 2: $|S_2|>0$ and $p_N\geq\frac{\epsilon}{2(M-N)}$. We have

$$\begin{aligned}
\min\left\{\frac{\epsilon}{2(M-N)},\min_{i\in S_3} p_i\right\} &\geq \min\left\{\frac{\epsilon}{2(M-N)},\min\left\{p_N,\frac{\epsilon}{2(M-N)}\right\}\right\}\\
&\geq \frac{\epsilon}{2(M-N)},
\end{aligned}$$

and

$$\begin{aligned}\max\left\{\max_{i\in S_2} p_i, \frac{\epsilon}{2(M-N)}\right\} &\leq \frac{\epsilon}{2(M-N)} \\ &\leq \min\left\{\frac{\epsilon}{2(M-N)}, \min_{i\in S_3} p_i\right\}.\end{aligned}$$

Case 3: $|S_2| > 0$ and $p_N < \frac{\epsilon}{2(M-N)}$. Then $|S_2| = M - N$. The R.H.S. of (34) becomes

$$\sum_{i\in S_2} f\left(\frac{\epsilon}{2(M-N)}, -\frac{\epsilon}{2(M-N)} + p_i + d_i\right) + \sum_{i\in S_3} f(p_i, d_i) - \sum_{i\in S_2} f\left(0, p_i\right).$$

We have

$$\begin{aligned}\min\left\{\frac{\epsilon}{2(M-N)}, \min_{i\in S_3} p_i\right\} &\geq \min\left\{\frac{\epsilon}{2(M-N)}, \min\left\{p_N, \frac{\epsilon}{2(M-N)}\right\}\right\} \\ &= p_N \\ &\geq p_{N+1} \\ &= \max_{i\in S_2}\{p_i\}.\end{aligned}$$

Therefore, (8) is checked in all possible cases and we can apply Lemma 5 to (34) and show that

$$\sum_{i:d_i>0} f(p_i, d_i) \leq (M-N) f\left(0, \frac{\epsilon}{2(M-N)}\right). \tag{35}$$

By putting (33) and (35) into (30), we have

$$\begin{aligned}H(\mathcal{Q}) - H(\mathcal{P}) &\leq Nf\left(a, -\frac{a\epsilon}{2}\right) + (M-N) f\left(0, \frac{\epsilon}{2(M-N)}\right) \\ &= \max_{\mathcal{Q}'\in\Gamma_M} (H(\mathcal{Q}') - H(\gamma)),\end{aligned}$$

where the last equality follows from (27).

In the previous part of this proof, we have assumed $V(\mathcal{P}, \mathcal{Q}) = \epsilon$. If $V(\mathcal{P}, \mathcal{Q}) = \eta < \epsilon$, the previous part of this proof tells

$$\begin{aligned}H(\mathcal{Q}) - H(\mathcal{P}) &\leq Nf\left(a, -\frac{a\eta}{2}\right) + (M-N) f\left(0, \frac{\eta}{2(M-N)}\right) \\ &= \max_{\mathcal{Q}''\in\Gamma_M} (H(\mathcal{Q}'') - H(\gamma)),\end{aligned}$$

where $\mathcal{Q}''$ is obtained from Theorem 6 subject to $V(\gamma, \mathcal{Q}'') \leq \eta$. By Corollary 7, it is readily seen that

$$\begin{aligned}H(\mathcal{Q}) - H(\mathcal{P}) &\leq \max_{\mathcal{Q}''\in\Gamma_M} (H(\mathcal{Q}'') - H(\gamma)) \\ &\leq \max_{\mathcal{Q}'\in\Gamma_M} (H(\mathcal{Q}') - H(\gamma)).\end{aligned}$$

The theorem is proved when we can find an upper bound on $H(\mathcal{P}) - H(\mathcal{Q})$ for any $\mathcal{P} \in \Gamma_L$ and $\mathcal{Q} \in \Gamma_M$ with $M \geq L$ such that $V(\mathcal{P}, \mathcal{Q}) \leq \epsilon$. By putting $\mathcal{P} = \mathcal{Q}''$ and $\epsilon = 2\sum_{i=L+1}^{M} q_i''$ into Theorem 8, we obtain $\mathcal{Q}^*$ where $H(\mathcal{Q}^*) < H(\mathcal{Q})$. Then $\mathcal{Q}^* = \{q_1 + \sum_{i=L+1}^{M} q_i, q_2, q_3, \ldots, q_L\}$ has at most L positive probability masses and the variational distance

$$\begin{aligned} V(\mathcal{P}, \mathcal{Q}^*) &= \sum_{i=1}^{L} |p_i - q_i^*| \\ &= |p_1 - q_1^*| + \sum_{i=2}^{L} |p_i - q_i^*| \\ &\leq |p_1 - q_1| + \sum_{i=L+1}^{M} q_i + \sum_{i=2}^{L} |p_i - q_i| \\ &= V(\mathcal{P}, \mathcal{Q}) \\ &\leq \epsilon. \end{aligned}$$

Since $\mathcal{P}$ and $\mathcal{Q}^*$ have the same number of positive probability masses, we can apply the results in the previous part and see that

$$\begin{aligned} \max_{\mathcal{P}' \in \Gamma_L : V(\mathcal{P}', \gamma) \leq \epsilon} (H(\mathcal{P}') - H(\gamma)) &\geq H(\mathcal{P}) - H(\mathcal{Q}^*) \\ &\geq H(\mathcal{P}) - H(\mathcal{Q}), \end{aligned}$$

where the last inequality follows from $H(\mathcal{Q}^*) < H(\mathcal{Q})$. Since $M \geq L$, it is readily checked that

$$\max_{\mathcal{P}' \in \Gamma_L : V(\mathcal{P}', \gamma) \leq \epsilon} (H(\mathcal{P}') - H(\gamma)) \leq \max_{\mathcal{Q}' \in \Gamma_M : V(\mathcal{Q}', \gamma) \leq \epsilon} (H(\mathcal{Q}') - H(\gamma))$$

As a whole, for all $\mathcal{P}$ and $\mathcal{Q}$,

$$\begin{aligned} |H(\mathcal{Q}) - H(\mathcal{P})| &\leq \max_{\mathcal{Q}' \in \Gamma_M : V(\mathcal{Q}', \gamma)} (H(\mathcal{Q}') - H(\gamma)) \\ &= N f\left(a, -\frac{a\epsilon}{2}\right) + (M - N) f\left(0, \frac{\epsilon}{2(M-N)}\right) \\ &= \begin{cases} H\left(\{\frac{\epsilon}{2}, 1 - \frac{\epsilon}{2}\}\right) + \frac{\epsilon}{2} \log \frac{M-N}{N} & 0 < \frac{\epsilon}{2} < \frac{M-N}{M} \\ \log M - \log N & \frac{\epsilon}{2} \geq \frac{M-N}{M}. \end{cases} \end{aligned}$$

■

When $a = 1$, we have the following important special case of Theorem 10.

Theorem 11: If $\mathcal{P} \in \Gamma_L$ and $\mathcal{Q} \in \Gamma_M$ with $M \geq L$ being two probability distributions such that $V(\mathcal{P}, \mathcal{Q}) \leq \epsilon$, then

$$\begin{aligned} & |H(\mathcal{Q}) - H(\mathcal{P})| \\ \leq & \begin{cases} H\left(\{\frac{\epsilon}{2}, 1-\frac{\epsilon}{2}\}\right) + \frac{\epsilon}{2}\log(M-1) & 0 < \frac{\epsilon}{2} < \frac{M-1}{M} \\ \log M & \frac{\epsilon}{2} \geq \frac{M-1}{M}. \end{cases} \end{aligned}$$

Thus for any fixed finite L and finite M, we have

$$\begin{aligned} 0 &\leq \lim_{\epsilon \to 0} \sup_{\mathcal{Q}} |H(\mathcal{Q}) - H(\mathcal{P})| \\ &\leq \lim_{\epsilon \to 0} \left[H\left(\left\{\frac{\epsilon}{2}, 1-\frac{\epsilon}{2}\right\}\right) + \frac{\epsilon}{2}\log(\max(L, M) - 1)\right] \\ &= 0. \end{aligned}$$

This can be interpreted as the Shannon entropy being continuous when the alphabet size is bounded and known.

We now compare the bound given in Theorem 11 with similar results in some literatures. For $\mathcal{P} \in \Gamma_L$ and $\mathcal{Q} \in \Gamma_M$, we let $N = \max(L, M) \geq 2$ and $V(\mathcal{P}, \mathcal{Q}) = \epsilon$ (c.f. (3)). Then (55) and (56) in [3] can be combined to become

$$|H(\mathcal{Q}) - H(\mathcal{P})| \leq g'(\epsilon, N),$$

where

$$g'(\epsilon, N) = \begin{cases} \epsilon \log N - \epsilon \log \epsilon & \epsilon \leq \frac{1}{3} \\ \epsilon(1 + \log N) - \epsilon \log \epsilon & \epsilon > \frac{1}{3}. \end{cases}$$

By Lemma 2.7 in [2], the bound $g'(\epsilon, N)$ can be improved to become

$$g(\epsilon, N) = \begin{cases} \epsilon \log N - \epsilon \log \epsilon & \epsilon \leq \frac{1}{2} \\ \epsilon(1 + \log N) - \epsilon \log \epsilon & \epsilon > \frac{1}{2}. \end{cases}$$

In order to compare $g(\epsilon, N)$ with the upper bound in Theorem 11, we consider two cases. We first consider that $0 < \epsilon \leq \frac{2(N-1)}{N}$ and let

$$\phi(\epsilon, N) = g(\epsilon, N) - H\left(\frac{\epsilon}{2}, 1-\frac{\epsilon}{2}\right) - \frac{\epsilon}{2}\log(N-1).$$

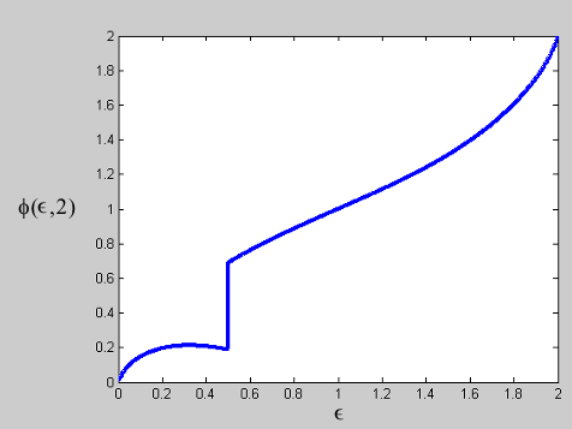

Fig. 6. A plotting of $\phi(\epsilon, 2)$ where the logarithms in ϕ are in base 2

Then our bound is tighter than $g(\epsilon, N)$ if we can show that $\phi \geq 0$ for all N and $0 < \epsilon < \frac{2(N-1)}{N} \leq 2$. Note that

$$\begin{aligned}
\frac{d\phi(\epsilon, N)}{dN} &= \frac{\epsilon}{N \ln 2} - \frac{\epsilon}{2(N-1)\ln 2} \\
&= \frac{2N\epsilon - 2\epsilon - \epsilon N}{2N(N-1)\ln 2} \\
&= \frac{\epsilon(N-2)}{2N(N-1)\ln 2} \\
&\geq 0
\end{aligned}$$

for $N \geq 2$. Therefore,

$$\phi(\epsilon, N) \geq \phi(\epsilon, 2).$$

Fig. 6 shows that $\phi(\epsilon, 2) \geq 0$ for $0 < \epsilon \leq 2$.

On the other hand, for $1 \leq \frac{2(N-1)}{N} < \epsilon \leq 2$, we consider

$$\begin{aligned}
g(\epsilon, N) - \log N &= \epsilon(1 + \log N) - \epsilon \log \epsilon - \log N \\
&= \epsilon - \epsilon \log \epsilon + \epsilon \log N - \log N \\
&\geq \epsilon - \epsilon \log \epsilon \\
&\geq \epsilon - \epsilon \log 2 \\
&= 0.
\end{aligned}$$

Finally, we note that the condition $V(\mathcal{P},\mathcal{Q}) = \epsilon$ in [3] is a special case of the condition $V(\mathcal{P},\mathcal{Q}) \leq \epsilon$ used in Theorem 11. The above calculations, however, show that our bounds are still smaller than $g(\epsilon, N)$. Therefore, some tighter bounds are obtained in this paper.

Note that the condition $V(\mathcal{P},\mathcal{Q}) = \epsilon$ in [3] is a special case of the condition $V(\mathcal{P},\mathcal{Q}) \leq \epsilon$ used in Theorem 11. However, our bounds are still smaller than $g(\epsilon, N)$. Therefore, some tighter bounds are shown in this paper. The bound given in Theorem 11 is, in fact, the tightest. Let

$$\mathcal{P} = \{1, 0, 0, \ldots, 0\}$$

and

$$\mathcal{Q} = \begin{cases} \left\{1 - \frac{\epsilon}{2}, \frac{\epsilon}{2(M-1)}, \ldots, \frac{\epsilon}{2(M-1)}\right\} & 0 < \frac{\epsilon}{2} < \frac{M-1}{M} \\ \left\{\frac{1}{M}, \frac{1}{M}, \ldots, \frac{1}{M}\right\} & \frac{\epsilon}{2} \geq \frac{M-1}{M}. \end{cases}$$

We have $V(\mathcal{P},\mathcal{Q}) \leq \epsilon$ and $|H(\mathcal{Q}) - H(\mathcal{P})|$ attaining the upper bound in Theorem 11.

Furthermore, Theorem 11 can be generalized to describe some distributions which have not been normalized. This result was used in [6]. We extend the definition of entropy to distribution which is not normalized.

Definition 1: For an unnormalized distribution $\tilde{\mathcal{P}} = (\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_L)$ which can be normalized by a positive constant $\alpha \leq 1$ so that $(\alpha^{-1}\tilde{p}_1, \alpha^{-1}\tilde{p}_2, \ldots, \alpha^{-1}\tilde{p}_L) \in \Gamma_L$, let

$$H(\tilde{\mathcal{P}}) = -\sum_{i=1}^{L} \tilde{p}_i \log \tilde{p}_i.$$

Theorem 12: Let $\tilde{\mathcal{P}} = (\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_L)$ and $\tilde{\mathcal{Q}} = (\tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_M)$ be two unnormalized distributions which can be normalized by two positive constants $\alpha \leq 1$ and $\beta \leq 1$ so that $(\alpha^{-1}\tilde{p}_1, \alpha^{-1}\tilde{p}_2, \ldots, \alpha^{-1}\tilde{p}_L) \in \Gamma_L$ and $(\beta^{-1}\tilde{q}_1, \beta^{-1}\tilde{q}_2, \ldots, \beta^{-1}\tilde{q}_M) \in \Gamma_M$ with $M \geq L$. If

$$V(\tilde{\mathcal{P}}, \tilde{\mathcal{Q}}) \leq \epsilon,$$

then

$$|H(\tilde{\mathcal{Q}}) - H(\tilde{\mathcal{P}})| \leq \begin{cases} -\epsilon \log \epsilon + \epsilon \log M & \epsilon < 1 \\ \log M & \epsilon \geq 1. \end{cases}$$

Proof: For any $\tilde{\mathcal{P}} = (\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_L)$ and $\tilde{\mathcal{Q}} = (\tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_L)$ such that $V(\mathcal{P},\mathcal{Q}) \leq \epsilon$, let $\tilde{d}_i$'s be some real values such that

$$(\tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_M) = (\tilde{p}_1 + \tilde{d}_1, \tilde{p}_2 + \tilde{d}_2, \ldots, \tilde{p}_L + \tilde{d}_L, \tilde{d}_{L+1}, \ldots, \tilde{d}_M).$$

It is obvious that

$$|H(\tilde{\mathcal{Q}}) - H(\tilde{\mathcal{P}})| \leq |\log M - 0| = \log M.$$

Suppose $\epsilon < 1$ in the following. For $\tilde{d}_i < 0$, by Lemma 3,

$$\begin{aligned} -(\tilde{p}_i + \tilde{d}_i)\log(\tilde{p}_i + \tilde{d}_i) + \tilde{p}_i \log \tilde{p}_i &\leq -(1 + \tilde{d}_i)\log(1 + \tilde{d}_i) + 1 \log 1 \\ &= -(1 + \tilde{d}_i)\log(1 + \tilde{d}_i). \end{aligned} \tag{36}$$

We now consider two cases. If $-1 < \tilde{d}_i < -0.5$, let

$$d_i^* = 1 + \tilde{d}_i < 0.5 \leq |\tilde{d}_i|,$$

which gives $d_i^* > 0$. Then by (36)

$$\begin{aligned} -(\tilde{p}_i + \tilde{d}_i)\log(\tilde{p}_i + \tilde{d}_i) + \tilde{p}_i \log \tilde{p}_i &\leq -(1 + \tilde{d}_i)\log(1 + \tilde{d}_i) \\ &= -d_i^* \log d_i^*. \end{aligned}$$

If $-0.5 \leq \tilde{d}_i < 0$, let

$$d_i^* = -\tilde{d}_i = |\tilde{d}_i|,$$

which again gives $d_i^* > 0$. Then by

$$-(1 - x)\log(1 - x) \leq -x \log x$$

for $0 \leq x \leq 0.5$ and (36), we have

$$\begin{aligned} -(\tilde{p}_i + \tilde{d}_i)\log(\tilde{p}_i + \tilde{d}_i) + \tilde{p}_i \log \tilde{p}_i &\leq -(1 + \tilde{d}_i)\log(1 + \tilde{d}_i) \\ &\leq -(1 - d_i^*)\log(1 - d_i^*) \\ &\leq -d_i^* \log d_i^*. \end{aligned}$$

Furthermore, for $\tilde{d}_i \geq 0$, let $d_i^* = \tilde{d}_i$. Then we have

$$\begin{aligned} -(\tilde{p}_i + \tilde{d}_i)\log(\tilde{p}_i + \tilde{d}_i) + \tilde{p}_i \log \tilde{p}_i &= f(\tilde{p}_i, \tilde{d}_i) \\ &\leq f(0, \tilde{d}_i) \\ &= f(0, d_i^*) \\ &= -d_i^* \log d_i^*, \end{aligned}$$

where the inequality follows from Lemma 2. Therefore, $d_i^* \geq 0$ for all i and

$$\sum_i d_i^* \leq \sum_i |\tilde{d}_i| \leq \epsilon,$$

but

$$\begin{aligned}
|H(\tilde{\mathcal{Q}}) - H(\tilde{\mathcal{P}})| &= \left| -\sum_i (\tilde{p}_i + \tilde{d}_i) \log(\tilde{p}_i + \tilde{d}_i) + \sum_i \tilde{p}_i \log \tilde{p}_i \right| \\
&\leq \sum_i | - (\tilde{p}_i + \tilde{d}_i) \log(\tilde{p}_i + \tilde{d}_i) + \tilde{p}_i \log \tilde{p}_i| \\
&\leq \sum_i | - d_i^* \log d_i^*| \\
&= -\sum_i d_i^* \log d_i^* \\
&\leq -M \cdot \frac{\epsilon}{M} \log \frac{\epsilon}{M} \\
&= -\epsilon \log \epsilon + \epsilon \log M.
\end{aligned}$$

As a whole, we have

$$|H(\tilde{\mathcal{Q}}) - H(\tilde{\mathcal{P}})| \leq \begin{cases} -\epsilon \log \epsilon + \epsilon \log M & \epsilon < 1 \\ \log M & \epsilon \geq 1. \end{cases}$$

■

The bound given in Theorem 12 is, in fact, the tightest. Let

$$\tilde{\mathcal{P}} = \{\delta, 0, 0, \ldots, 0\}$$

and

$$\tilde{\mathcal{Q}} = \begin{cases} \left\{\frac{\epsilon}{M}, \frac{\epsilon}{M}, \ldots, \frac{\epsilon}{M}\right\} & 0 < \epsilon < 1 \\ \left\{\frac{1}{M}, \frac{1}{M}, \ldots, \frac{1}{M}\right\} & \epsilon \geq 1, \end{cases}$$

where $\delta \approx 0$. We have $V(\tilde{\mathcal{P}}, \tilde{\mathcal{Q}}) \leq \epsilon$ and $|H(\tilde{\mathcal{Q}}) - H(\tilde{\mathcal{P}})|$ attaining the upper bound in Theorem 12.

Note that Theorem 12 is the same as Lemma 2.7 in [2] for $\epsilon \leq 0.5$ because the proof of Lemma 2.7 in [2] has not used the fact that probability distributions are normalized although it is implicitly assumed. This explains why their bound is not tight and looks similar with Theorem 12. Moreover, Lemma 2.7 in [2] requires that $\epsilon \leq 0.5$ which is not required in Theorem 12.

In Theorem 11 and Theorem 12, upper bounds on $\sup_{\mathcal{Q}} |H(\mathcal{Q}) - H(\mathcal{P})|$ have been obtained. For the sake of completeness, a lower bound on $\sup_{\mathcal{Q}} |H(\mathcal{Q}) - H(\mathcal{P})|$ is given after the proof of the following theorem which is a refinement of Theorem 1.

Theorem 13: If $\mathcal{P} \in \Gamma_L$ and $\mathcal{Q} \in \Gamma_M$ with $M \geq L$ being two probability distributions such that

$$V(\mathcal{P}, \mathcal{Q}) \leq \epsilon,$$

then

$$\inf_{\mathcal{P}} \sup_{\mathcal{Q}} (H(\mathcal{Q}) - H(\mathcal{P})) = \begin{cases} H\left(\frac{\epsilon}{2}, 1 - \frac{\epsilon}{2}\right) + \frac{\epsilon}{2} \log\left(\frac{M}{L} - 1\right) & 0 < \frac{\epsilon}{2} < \frac{M-L}{M} \\ \log M - \log L & \frac{\epsilon}{2} \geq \frac{M-L}{M}. \end{cases} \tag{37}$$

Proof: It is not obvious that a lower bound on

$$\sup_{\mathcal{Q}} (H(\mathcal{Q}) - H(\mathcal{P})) \tag{38}$$

is given by taking $\mathcal{P}$ to be the uniform distribution, but it will be seen to be true. Let $\mathcal{P} \in \Gamma_L$ and $\mathcal{Q} \in \Gamma_M$ be any two distributions where $\mathcal{Q}$ is obtained from $\mathcal{P}$ according to Theorem 6 subject to $V(\mathcal{P}, \mathcal{Q}) \leq \epsilon$. We use the notations in (1) and (2) for $\mathcal{P}$ and $\mathcal{Q}$, respectively, and we assume that p_i's are sorted in descending order. Furthermore, let

$$\mathcal{P}' = \left(\frac{1}{L}, \frac{1}{L}, \ldots, \frac{1}{L}\right)$$

be the uniform distribution in Γ_L, and

$$\mathcal{Q}' = \left(\frac{1}{L} + d'_1, \frac{1}{L} + d'_2, \ldots, \frac{1}{L} + d'_L, d'_{L+1}, \ldots, d'_M\right)$$

be the distribution in Γ_M obtained by Theorem 6 subject to $V(\mathcal{P}', \mathcal{Q}') \leq \epsilon$. We will show that

$$\inf_{\mathcal{P}} \sup_{\mathcal{Q}: V(\mathcal{P}, \mathcal{Q}) \leq \epsilon} (H(\mathcal{Q}) - H(\mathcal{P})) = H(\mathcal{Q}') - H(\mathcal{P}') \tag{39}$$

by considering three cases.

We first consider the case that $V(\mathcal{P}, \mathcal{Q}) < \epsilon$. Let μ and ν satisfy (9) and (10) where $\mathcal{Q}$ is obtained from $\mathcal{P}$ in Theorem 6. If $\mu > \nu$, then $V(\mathcal{P}, \mathcal{Q}) = \epsilon$ from the construction of $\mathcal{Q}$. By contradiction, we have shown that

$$\mathcal{Q} = \left(\frac{1}{M}, \frac{1}{M}, \ldots, \frac{1}{M}\right).$$

Note that

$$\begin{aligned}
\epsilon &> V(\mathcal{P},\mathcal{Q}) \\
&= \sum_{i=1}^{L}\left|p_i-\frac{1}{M}\right|+(M-L)\cdot\frac{1}{M} \\
&= \sum_{i:p_i\geq\frac{1}{M}}\left(p_i-\frac{1}{M}\right)+\sum_{i:p_i<\frac{1}{M}}\left(\frac{1}{M}-p_i\right)+(M-L)\cdot\frac{1}{M} \\
&= \sum_{i:p_i^*\geq\frac{1}{M}}\left(p_i^*-\frac{1}{M}\right)-\sum_{i:p_i^*<\frac{1}{M}}\left(\frac{1}{M}-p_i^*\right)+2\cdot\sum_{i:p_i^*<\frac{1}{M}}\left(\frac{1}{M}-p_i^*\right)+1-\frac{L}{M} \\
&= \sum_{i=1}^{L}\left(p_i-\frac{1}{M}\right)+2\cdot\sum_{i:p_i<\frac{1}{M}}\left(\frac{1}{M}-p_i\right)+1-\frac{L}{M} \\
&\geq \left(1-\frac{L}{M}\right)+0+1-\frac{L}{M} \\
&= 2\cdot\left(1-\frac{L}{M}\right).
\end{aligned}$$

Then

$$V(\mathcal{P}',\mathcal{Q})=L\cdot\left(\frac{1}{L}-\frac{1}{M}\right)+(M-L)\cdot\frac{1}{M}=2\cdot\left(1-\frac{L}{M}\right)<\epsilon.$$

Therefore, $\log M = H(\mathcal{Q}) \leq H(\mathcal{Q}')$ so that $H(\mathcal{Q}') = \log M$ and hence $\mathcal{Q}'$ is the uniform distribution. Since $V(\mathcal{P}',\mathcal{Q}') = V(\mathcal{P}',\mathcal{Q}) < \epsilon$ and $H(\mathcal{P}') \geq H(\mathcal{P})$, (38) is minimized when $\mathcal{P}=\mathcal{P}'$ in this case.

Now we assume that $V(\mathcal{P},\mathcal{Q})=\epsilon$ and $V(\mathcal{P}',\mathcal{Q}')=\epsilon$. By construction of $\mathcal{Q}$,

$$d_1'=d_2'=\cdots=d_L'=-\frac{\epsilon}{2L}$$

and

$$d_{L+1}'=d_{L+2}'=\cdots=d_M'=\frac{\epsilon}{2(M-L)}.$$

Note that

$$\sum_{i:d_i>0}d_i=\sum_{i:d_i'>0}d_i'=\frac{\epsilon}{2}.$$

We claim that

$$\sum_{i:d_i>0}f(p_i,d_i)\geq\sum_{i:d_i'>0}f\left(0,d_i'\right). \tag{40}$$

Otherwise, consider the example in Fig. 2 and let $\mathcal{Q}' = \{\mu, \mu, \mu, p_4, p_5, p_6, p_7, \frac{\epsilon}{4}, \frac{\epsilon}{4}\}$. If (40) is false, then $H(\mathcal{Q}') \geq H(\mathcal{Q}^*)$ in Theorem 6 which causes contradiction. We now use Lemma 4 to show that

$$\sum_{i:d_i<0} f(p_i, d_i) - Lf\left(\frac{1}{L}, -\frac{\epsilon}{2L}\right) \tag{41}$$

is nonnegative. It is easily to check that $\mu \geq \frac{1}{L} - \frac{\epsilon}{2L}$. Assume $\mu \geq \frac{1}{L}$. We have

$$\min_{i:d_i<0}\{p_i + d_i\} = \mu \geq \frac{1}{L}.$$

Then, we can apply Lemma 4 to show that (41) is nonnegative. Assume $\mu \leq \frac{1}{L}$. Define two sets of positive integers

$$S_0 = \left\{i : d_i < 0 \text{ and } p_i > \frac{1}{L}\right\}$$

and

$$S_1 = \{i : d_i < 0 \text{ and } i \notin S_0\}.$$

Then

$$\begin{aligned}
&\sum_{i:d_i<0} f(p_i, d_i) - Lf\left(\frac{1}{L}, -\frac{\epsilon}{2L}\right) \\
&= \sum_{i\in S_0}\left(f(p_i, d_i) - f\left(\frac{1}{L}, -\frac{\epsilon}{2L}\right)\right) + \sum_{i\in S_1}\left(f(p_i, d_i) - f\left(\frac{1}{L}, -\frac{\epsilon}{2L}\right)\right) \\
&\quad -(L - |S_0| - |S_1|)f\left(\frac{1}{L}, -\frac{\epsilon}{2L}\right) \\
&= \sum_{i\in S_0}\left(f\left(p_i, -p_i + \frac{1}{L}\right) + f\left(\frac{1}{L}, -\frac{1}{L} + \mu\right) - f\left(\frac{1}{L}, -\frac{1}{L} + \mu\right) - f\left(\mu, -\mu + \frac{1}{L} - \frac{\epsilon}{2L}\right)\right) \\
&\quad + \sum_{i\in S_1}\left(f(p_i, d_i) - f\left(\frac{1}{L}, -\frac{1}{L} + p_i\right) - f\left(p_i, d_i\right) - f\left(\mu, -\mu + \frac{1}{L} - \frac{\epsilon}{2L}\right)\right) \\
&\quad -(L - |S_0| - |S_1|)f\left(\frac{1}{L}, -\frac{\epsilon}{2L}\right) \\
&= \sum_{i\in S_0} f\left(p_i, -p_i + \frac{1}{L}\right) - |S_0|f\left(\mu, -\mu + \frac{1}{L} - \frac{\epsilon}{2L}\right) \\
&\quad - \sum_{i\in S_1} f\left(\frac{1}{L}, -\frac{1}{L} + p_i\right) - |S_1|f\left(\mu, -\mu + \frac{1}{L} - \frac{\epsilon}{2L}\right) - (L - |S_0| - |S_1|)f\left(\frac{1}{L}, -\frac{\epsilon}{2L}\right)
\end{aligned} \tag{42}$$

Note that the canceled terms in the above calculation do not affect that (42) still satisfies the requirement in (5). Furthermore,

$$\min_{i\in S_0}\left\{p_i - p_i + \frac{1}{L}\right\} = \frac{1}{L} = \max\left\{\frac{1}{L}, \mu\right\}.$$

Then, we can apply Lemma 4 to show that (41) is nonnegative. Hence,

$$\sum_{i:d_i<0} f(p_i, d_i) - \sum_{i:d_i'<0} f\left(\frac{1}{L}, d_i'\right) = \sum_{i:d_i<0} f(p_i, d_i) - Lf\left(\frac{1}{L}, -\frac{\epsilon}{2L}\right) \geq 0.$$

Therefore,

$$\sum_{i:d_i<0} f(p_i, d_i) \geq \sum_{i:d_i'<0} f\left(\frac{1}{L}, d_i'\right).$$

Together with (40), we can conclude that

$$\begin{aligned} H(\mathcal{Q}) - H(\mathcal{P}) &= \sum_i f(p_i, d_i) \\ &\geq \sum_{i:d_i'>0} f\left(0, d_i'\right) + \sum_{i:d_i'<0} f\left(\frac{1}{L}, d_i'\right) \\ &= H(\mathcal{Q}') - H(\mathcal{P}'). \end{aligned}$$

Finally we assume that $V(\mathcal{P}, \mathcal{Q}) = \epsilon$ but $V(\mathcal{P}', \mathcal{Q}') = \delta < \epsilon$. This means that $\mathcal{Q}'$ is the uniform distribution. Then a $\mathcal{Q}^*$ can be obtained from $\mathcal{P}$ according to Theorem 6 subject to $V(\mathcal{P}, \mathcal{Q}^*) = \delta$. Therefore, we have

$$\begin{aligned} H(\mathcal{Q}) - H(\mathcal{P}) &\geq H(\mathcal{Q}^*) - H(\mathcal{P}) \\ &\geq H(\mathcal{Q}') - H(\mathcal{P}'), \end{aligned}$$

where the first inequality follows from $\delta < \epsilon$ and Corollary 7 and the second inequality follows from the result in the last paragraph.

Thus we have proved (39) in all possible cases. The value of d_i' and $\mathcal{Q}'$ can be obtained from Theorem 6. Therefore,

$$\begin{aligned} \inf_{\mathcal{P}} \sup_{\mathcal{Q}} \left(H(\mathcal{Q}) - H(\mathcal{P})\right) &= H(\mathcal{Q}') - H(\mathcal{P}') \\ &= \begin{cases} H\left(\frac{\epsilon}{2}, 1-\frac{\epsilon}{2}\right) + \frac{\epsilon}{2}\log\left(\frac{M}{L} - 1\right) & 0 < \frac{\epsilon}{2} < \frac{M-L}{M} \\ \log M - \log L & \frac{\epsilon}{2} \geq \frac{M-L}{M}. \end{cases} \end{aligned}$$

■

Note that

$$\inf_{\mathcal{P}} \sup_{\mathcal{Q}} |H(\mathcal{Q}) - H(\mathcal{P})| \geq \inf_{\mathcal{P}} \sup_{\mathcal{Q}} (H(\mathcal{Q}) - H(\mathcal{P})) \tag{43}$$

$$= \begin{cases} H\left(\frac{\epsilon}{2}, 1-\frac{\epsilon}{2}\right) + \frac{\epsilon}{2}\log\left(\frac{M}{L}-1\right) & 0 < \frac{\epsilon}{2} < \frac{M-L}{M} \\ \log M - \log L & \frac{\epsilon}{2} \geq \frac{M-L}{M} \end{cases} \tag{44}$$

from Theorem 13. Although the bound in (43) is not tight, Theorem 13 is strong enough to subsume Theorem 1 because for any fixed $\epsilon > 0$, the R.H.S. of (44) tends to infinity as $M \to \infty$.

III. Conclusion

We have introduced the way to find the distribution which attains the minimum or the maximum entropy within a given variational distance from a given probability distribution. For any two probability distributions, we have obtained the tightest upper bound on the difference of their entropies in terms of their alphabet sizes and variational distance. The lower bound of the difference has also been obtained. These bounds have related the continuity/discontinuity of entropy and the alphabet size of a distribution. The applications of these results will be shown in Part II of this paper.

Acknowledgment

The author would like to thank Sergio Verdú for his valuable comments.

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, New York: Wiley-Interscience, 1991.

[2] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems,* Academic Press, New York, 1981.

[3] J. Naudts, "Continuity of a class of entropies and relative entropies," Rev. Math. Phys., 16:809-822, 2004.

[4] S.-W. Ho and R. W. Yeung, "On the Discontinuity of the Shannon Information Measures," in *Proc. 2005 IEEE Int. Symposium Inform. Theory (ISIT 2005)*, Adelaide, Australia, Sept. 4-9, 2005.

[5] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications,* Academic Press, New York, 1979.

[6] S.-W. Ho and R. W. Yeung, "On Information Divergence Measures and a Unified Typicality," in *Proc. 2006 IEEE Int. Symposium Inform. Theory (ISIT 2006)*, Seattle, United States, July 9-14, 2006.